P170M109 Computational Intelligence and Decision Making

Supervised learning (part 1)

ktu
1922

2022

K-nearest neighbors (KNN)

classification

Input: labeled data, K (number of nearest neighbors), query example

1) For each example in dataset, calculate distance to query example

2) Select K examples with the smallest distances

3) Return **mode** of K labels

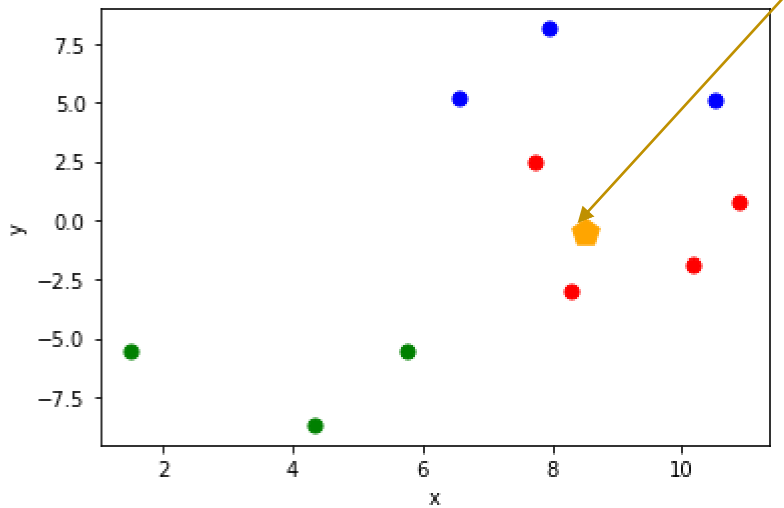
Output: predicted **label** for query example

K-nearest neighbors (KNN)

classification

Query (8.5; -0.5)

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$




```
colors = {0:'red', 1:'blue', 2:'green'}
```

	x	y	label	distance
0	7.72	2.44	0	3.04
1	1.50	-5.54	2	8.63
2	**8.29**	**-3.02**	**0**	**2.53**
3	10.52	5.13	1	5.98
4	6.56	5.17	1	6.00
5	**10.89**	**0.72**	**0**	**2.68**
6	4.35	-8.70	2	9.19
7	7.94	8.16	1	8.68
8	**10.17**	**-1.92**	**0**	**2.20**
9	5.76	-5.56	2	5.76

K = 3

Indices of selected neighbors: 2; 5; 8

Labels of selected neighbors: 0; 0; 0

Predicted label: 0 (red)

K-nearest neighbors (KNN)

regression

Input: data (with known values), K (number of nearest neighbors), query example

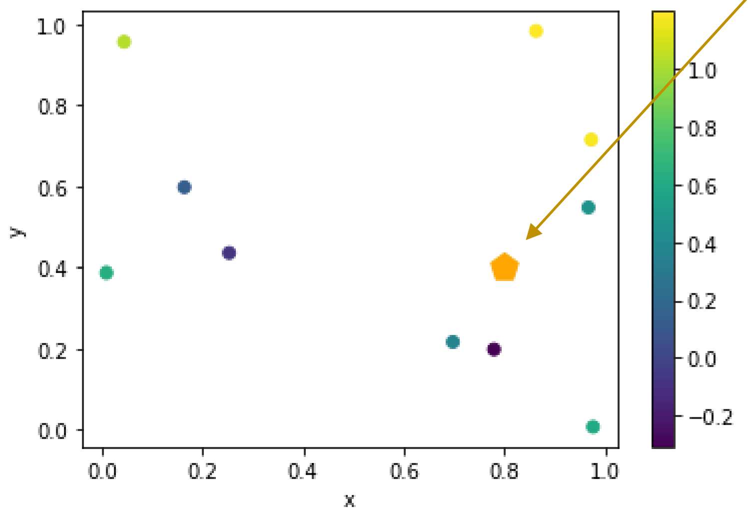
1) For each example in dataset, calculate distance to query example
2) Select K examples with the smallest distances
3) Return **mean** of K values

Output: predicted **value** for query example

K-nearest neighbors (KNN)

regression

Query
(0.8; 0.4)



$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

	x	y	value	distance
0	**0.97**	**0.55**	**0.45**	**0.22**
1	0.97	0.71	1.18	0.36
2	**0.70**	**0.22**	**0.37**	**0.21**
3	0.98	0.01	0.59	0.43
4	0.25	0.43	-0.08	0.55
5	**0.78**	**0.20**	**-0.31**	**0.20**
6	0.86	0.98	1.20	0.59
7	016	0.60	0.14	0.67
8	0.01	0.39	0.62	0.79
9	004	0.96	1.03	0.94

K = 3

Indices of selected neighbors: 0; 2; 5

Predicted value: $\frac{0.45 + 0.37 - 0.31}{3} = 0.17$

Example No. 2 KNN for continuous data

Code example:

- KNN_continuous.ipynb
- mixedDataExample.tsv

To do:

- Split initial dataset to train test and test dataset
- Calculate prediction accuracy

	LotArea	OverallQual	YearBuilt	SalePrice
0	8450	7	2003	208500
1	9600	6	1976	181500
2	11250	7	2001	223500
3	9550	7	1915	140000
4	14260	8	2000	250000

```
Predicted price for
   LotArea  OverallQual  YearBuilt
0     8500            5       2000
predicted price: 145000.00
```

Example No. 2 KNN for continuous and categorical data

Code example:

- KNN_continuous_categorical.ipynb
- mixedDataExample.tsv

To do:

- Split initial dataset to train test and test dataset
- Calculate prediction accuracy

	LotArea	OverallQual	YearBuilt	RoofStyle	CentralAir	SalePrice
0	8450	7	2003	Gable	Y	208500
1	9600	6	1976	Gable	Y	181500
2	11250	7	2001	Gable	Y	223500
3	9550	7	1915	Gable	Y	140000
4	14260	8	2000	Gable	Y	250000

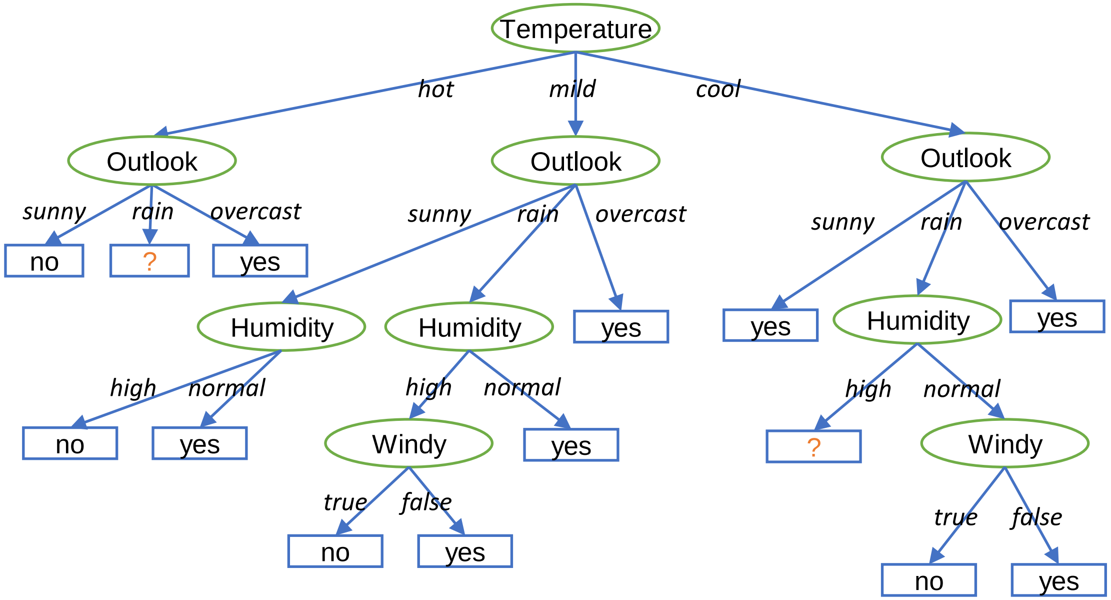
```
Predicted price for
   LotArea  OverallQual  YearBuilt  ...  roof_Hip  roof_Mansard  roof_Shed
0    18500            5       1960  ...         0             0          0

[1 rows x 10 columns]
predicted price: 163900.00
```

Decision Tree

A Decision Tree is a tree with nodes representing deterministic decisions based on variables and edges representing path to next node or a leaf node based on the decision

Temperature	Outlook	Humidity	Windy	Play?
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	overcast	high	false	yes
cold	rain	normal	false	yes
cold	overcast	normal	true	yes
mild	sunny	high	false	no
cold	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no



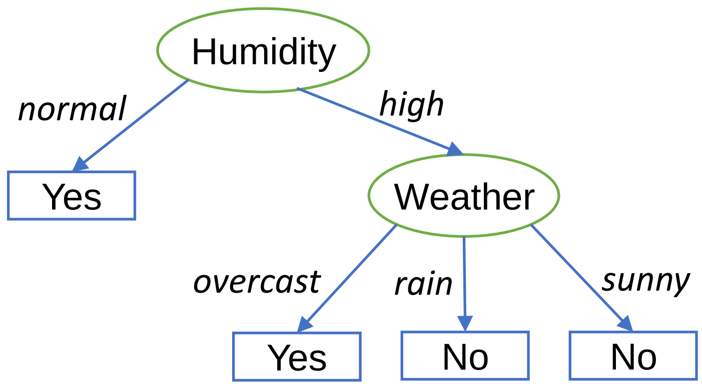
Decision Tree

Straightforward construction of *decision tree* leads that tree size **exponentially** depends on input data.

At the worst-case scenario n categorical features with **m** possible (*without taking in to account continuous variables!*) values each will have $O(m^n)$, for binary features, when $m = 2$ it is $\boldsymbol{O(2^n)}$.

A Decision Tree is modelled on a **simple series of questions** that lead serially to an answer that best fits the data used in training.

Temperature	Outlook	Humidity	Windy	Play?
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	overcast	high	false	yes
cold	rain	normal	false	yes
cold	overcast	normal	true	yes
mild	sunny	high	false	no
cold	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no



Decision tree - Algorithm

Step 1: Build the root with variables of most importance

Step2: Build a decision of **highest information split**

Step3: **Recursively** construct the nodes and decision using 1 and 2 step until no information can be split on the edge node

Highest information split

The exact decision (attribute selection) at construction of decision tree is generally performed using **Information gain** or **Gini impurity** criterion.

- **Information gain** is used if variables are categorical, i.e., if values fall into classes or categories and do not have a logical order (i.e. types of fruits)
- **Gini impurity** is used if the values are continuous, i.e., the values are numerical, for example, age of a person.

Entropy

Less impure node requires less information to describe it. And, more impure node requires more information.

Using information theory, we estimate the amount of information contained within each variable. A key measure in information theory is **entropy**.

$$H = -\sum_{i=1}^{n} (p_i \log_2 p_i) \Rightarrow$$

$\boldsymbol{p_i}$ is probability of occurrence of $\boldsymbol{i}$ -th possible value
$\boldsymbol{n}$ is the number of values
$\boldsymbol{H}$ is measure of entropy

Dice with possible values of 1–6

$$H = -\sum_{i=1}^{6} \left(\frac{1}{6}\log\frac{1}{6}\right) \approx 2.58$$

Flip coin

$$H = -\sum_{i=1}^{2} \left(\frac{1}{2}\log\frac{1}{2}\right) \approx 1$$

Entropy

Variable 1	Variable 2	Outcome
3	5	Stop
7	6	Continue
3	3	Stop
4	8	Continue
3	9	Continue
6	5	Stop
5	8	Continue
6	4	Continue

$$\boldsymbol{H(outcome) = -\left(\frac{3}{8}\log\frac{3}{8}\right) - \left(\frac{5}{8}\log\frac{5}{8}\right) \approx 0.954}$$

Decision on Variable 1: >**4 (greater than 4)**

Decision on Variable 2: >**6 (greater than 4)**

*based on variable average **

Entropy

$$\boldsymbol{H(outcome) = -\left(\frac{3}{8}\log\frac{3}{8}\right) - \left(\frac{5}{8}\log\frac{5}{8}\right) \approx 0.954}$$

Entropy before taking decision

$$\boldsymbol{H(>4, variable1) = -\left(\tfrac{3}{4}\log\tfrac{3}{4}\right) - \left(\tfrac{1}{4}\log\tfrac{1}{4}\right) \approx 0.81}$$

$$\boldsymbol{H(\leq 4, variable1) = -\left(\tfrac{2}{4}\log\tfrac{2}{4}\right) - \left(\tfrac{2}{4}\log\tfrac{2}{4}\right) \approx 1}$$

$$\boldsymbol{H(>6, variable2) = -\left(\tfrac{3}{3}\log\tfrac{3}{3}\right) - \left(\tfrac{0}{3}\log\tfrac{0}{3}\right) = 0}$$

$$\boldsymbol{H(\leq 6, variable2) = -\left(\tfrac{2}{5}\log\tfrac{2}{5}\right) - \left(\tfrac{3}{5}\log\tfrac{3}{5}\right) = 0.97}$$

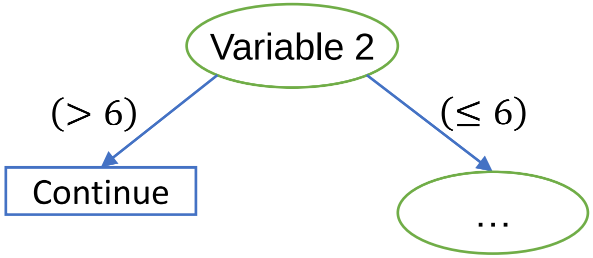
$$\boldsymbol{H(outcome, variable1)} = -p_{>4}H(>4) - p_{\leq 4}H(\leq 4) = \boldsymbol{\left(\tfrac{4}{8}\right) * 0.81 + \left(\tfrac{4}{8}\right) * 1 = 0.9}$$

$$\boldsymbol{H(outcome, variable2)} = -p_{>6}H(>6) - p_{\leq 6}H(\leq 6) = \boldsymbol{\left(\frac{3}{8}\right) * 0 + \left(\frac{5}{8}\right) * 0.97 = 0.61}$$

Entropy after decision

$$\boldsymbol{IG = H(outcome) - H(outcome, variable1) = 0.954 - 0.9 = 0.054}$$

$$\boldsymbol{IG = H(outcome) - H(outcome, variable2) = 0.954 - 0.61 = 0.344}$$



Exercise 2

Temperature	Outlook	Humidity	Windy	Play?
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	overcast	high	false	yes
cold	rain	normal	false	yes
cold	overcast	normal	true	yes
mild	sunny	high	false	no
cold	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no

$$\boldsymbol{IG = H(play) - H(play, temperature) = \_\_\_}$$

$$\boldsymbol{IG = H(play) - H(play, outlook) = \_\_\_}$$

$$\boldsymbol{IG = H(play) - H(play, windy) = \_\_\_}$$

8 minutes

Advantages

- Easy to Understand
- Useful in Data exploration
- Less data cleaning required
- Data type is not a constraint

Disadvantage

- Greedy solution
- Prone to overfitting

Questions?